Computing and Informatics, Vol. 31, 2012, 245–270

STUDY AND COMPARISON OF RULE-BASED AND STATISTICAL CATALAN-SPANISH MACHINE TRANSLATION SYSTEMS

Marta R. Costa-Jussà

Barcelona Media Innovation Center
Av. Diagonal 177, 08018 Barcelona, Spain
e-mail: marta.ruiz@barcelonamedia.org

Mireia Farrús

Universitat Oberta de Catalunya
Av. Tibidabo, 47. 08035 Barcelona, Spain
e-mail: mfarrus@uoc.edu

José B. Mariño, José A. R. Fonollosa

Universitat Politècnica de Catalunya, TALP Research Center
Jordi Girona 1-3, 08034 Barcelona, Spain
e-mail: {jose.marino, jose.fonollosa}@upc.edu

Communicated by Eva Hajičová

Abstract. Machine translation systems can be classified into rule-based and corpus-based approaches, in terms of their core methodology. Since both paradigms have been largely used during the last years, one of the aims in the research community is to know how these systems differ in terms of translation quality. To this end, this paper reports a study and comparison of several specific Catalan-Spanish machine translation systems: two rule-based and two corpus-based (particularly, statistical-based) systems, all of them freely available on the web. The translation quality analysis is performed under two different domains: journalistic and medical. The systems are evaluated by using standard automatic measures, as well as by native human evaluators. In addition to these traditional evaluation procedures, this paper

reports a novel linguistic evaluation, which provides information about the errors encountered at the orthographic, morphological, lexical, semantic and syntactic levels. Results show that while rule-based systems provide a better performance at orthographic and morphological levels, statistical systems tend to commit less semantic errors. Furthermore, results show all the evaluations performed are characterised by some degree of correlation, and human evaluators tend to be specially critical with semantic and syntactic errors.

Keywords: Rule-based machine translation, statistical machine translation, Catalan, Spanish

Mathematics Subject Classification 2010: 68, 68T50

1 INTRODUCTION

Machine Translation (MT) is a subfield of computational linguistics that investigates the use of computer software to translate text from one given source language to another target language. Since natural languages are highly complex, MT becomes a difficult task. Many words have multiple meanings, sentences may have various readings, and certain grammatical relations in one language might not exist in another language. Moreover, there are non-linguistic factors such as the need of having a world knowledge to perform a translation. In order to face the MT challenge, many dependencies have to be taken into account. These are often weak and vague, which makes it rarely possible to describe simple and relevant rules that hold without exception for different language pairs.

1.1 Motivation of MT

At the end of the 19^{th} century, L. L. Zamenhof proposed Esperanto, which was intended as a global language to be spoken and understood by everyone. The inventor was hoping that a common language could resolve global problems that lead to conflict. Esperanto as a planned language might have had some success, but nowadays, the Information Society is and will continue to be multilingual. While surfing the Internet, for instance, sometimes we come across languages and characters we don't understand. In this context, translation is the bottleneck of the pretended information globalisation. MT may also be used in text or e-mail translation to the desired spoken language. Not only has MT created great expectations, but it also is the only solution to some situations. Europe, for instance, with more than twenty official languages – and the number is continuously increasing – is lost in translation without MT. The European Union institutions currently employ around 2 000 written-text translators and they also need 80 interpreters per language and per day. Most translations regard to administrative reports, instruction manuals and other

documents that have neither cultural nor literary value. In general, the demand for translations is increasing more quickly than the capacity of translators, apart from the problem of the lack of qualified candidates for some languages, which makes MT even more necessary.

1.2 MT Classified by Its Core Methodology

MT systems can be classified according to their core methodology. Under this classification, two main paradigms can be found: the rule-based approach and the corpus-based approach. In the rule-based approach, human experts specify a set of rules to describe the translation process, so that an enormous amount of input from human experts is required [13, 1].

On the other hand, under the corpus-based approach the knowledge is automatically extracted by analysing translation examples from a parallel corpus built by human experts. The advantage is that, once the required techniques have been developed for a given language pair, MT systems should – theoretically – be quickly developed for new language pairs using provided training data.

Within the corpus-based paradigm, two other approaches can be further distinguished: example-based and statistical-based. Example-based machine translation (EBMT) makes use of previously seen examples in parallel corpora. EBMT is often characterized by the use of a bilingual corpus with parallel texts as its main knowledge base, at run-time. It is essentially a translation by analogy and it can be viewed as an implementation of case-based reasoning approach of machine learning. In statistical machine translation (SMT), parallel examples are used to train a statistical translation model. Thus, it relies on statistical parameters and a set of translation and language models, among other data-driven features. This approach worked initially on a word-by-word basis. However, current systems attempt to introduce a certain degree of linguistic analysis into the SMT approach. Nowadays, the most widely used MT systems use the rule-based and the statistical approaches. Moreover, there have been several research works which combine both methodologies [20]. Our study is intended to reinforce the system combination research works by further analysing both the structure of the two methodologies, and the specific type of errors which they tend to commit.

1.3 Structure of the Paper

Increasing computational power picked the current interest in MT. As a consequence, available machine translation systems in the web are becoming more and more popular. In the specific case of Catalan-Spanish translation, the available systems use two MT core methodologies: rule-based and statistical MT systems. Section 2 provides an extensive comparison of the rule-based and the statistical-based systems at the level of core methodology. A structural comparison of these methodologies is made in Section 3, by describing their challenges and general advantages and disadvantages. Section 4 reports a brief description of four Catalan-Spanish MT freely-available

systems: Apertium and Translendium as rule-based systems and Google and UPC as statistical systems. Section 5 describes the experimental framework used to compare the cited systems. Sections 6, 7 and 8 provide an exhaustive comparison by using automatic, human and linguistic evaluation, respectively. Finally, Section 9 presents the conclusions.

2 RULE-BASED VS. STATISTICAL-BASED MACHINE TRANSLATION

Rule-based machine translation (RBMT) systems were the first commercial machine translation systems. Much more complex than translating word to word, these systems develop linguistic rules that allow the words to be put in different places, to have different meaning depending on context, etc. The Georgetown-IBM experiment in 1954 was one of the first rule-based machine translation systems and Systran was one of the first companies to develop RBMT systems.

RBMT methodology applies a set of linguistic rules in three different phases: analysis, transfer and generation. Therefore, a rule-based system requires: syntax analysis, semantic analysis, syntax generation and semantic generation. Speaking in general terms, RBMT generates the target text given a source text following the steps shown in Figure 1.

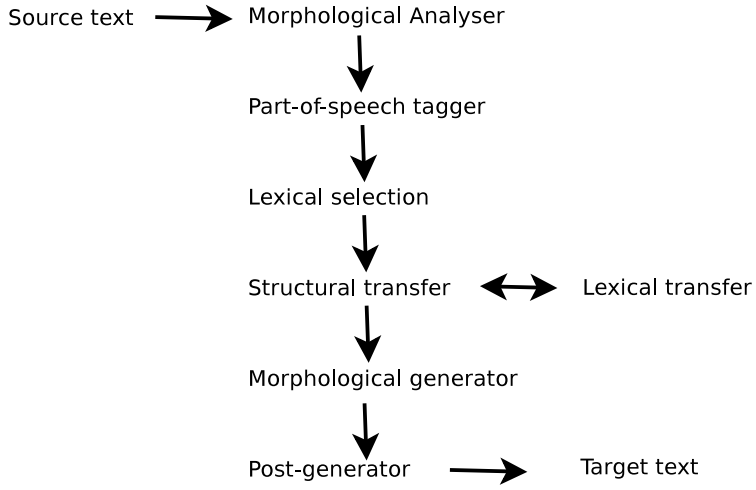

Fig. 1. Architecture of the RBMT approach

Given a source text, the first step is to segment it, for instance, by expanding elisions or marking set phrases. These segments are then looked up in a dictionary. This search returns the base form and tags for all matches (morphological analyser). Afterwards, the task is to resolve ambiguous segments, i.e. source terms that have

more than one match, by choosing only one (part of speech tagger). Additionally, a RBMT system may add a lexical selection to choose between alternative meanings. After the module taking care of the lexical selection, two modules follow, namely the structural and the lexical transfers. The former consists of looking up disambiguated source-language base work to find the target-language equivalent. The latter consists in:

1. flagging grammatical divergences between source language and target language, e.g. gender or number agreement;
2. creating a sequence of chunks;
3. reordering or modifying chunk sequences; and
4. substituting fully-tagged target-language forms into the chunks.

Then, tags are used to deliver the correct target language surface form (morphological generator). Finally, the last step is to make any necessary orthographic change (post-generator).

One of the main problems of translation is choosing the correct meaning, which involves a classification or disambiguation problem. In order to improve the accuracy, it is possible to apply a method to disambiguate meanings of a single word. Machine learning techniques automatically extract the context features that are useful for disambiguating a word.

RBMT systems have a big drawback: the construction of such systems demands a great amount of time and linguistic resources; as a result, it is very expensive. Moreover, in order to improve the quality of a RBMT it is necessary to modify rules, which requires more linguistic knowledge. Modification of one rule cannot guarantee that the overall accuracy will be better. However, using rule-based methodology may be the only way to build an MT system, given that SMT requires massive amounts of sentence-aligned parallel text (*is there such a resource for Icelandic?*). Additionally, the use of linguists may be a good choice. RBMT may use linguistic data elicited by speakers without access to existing machine-readable resources and it is more transparent: errors are easier to diagnose and debug.

Statistical Machine Translation (SMT), which started with the CANDIDE system [3], is, at its most basic, a more complicated form of word translation, where statistical weights are used to decide the most likely translation of a word. Modern SMT systems are phrase-based rather than word-based, and assemble translations using the overlap in phrases.

The main goal of SMT is the translation of a text given in some source language into a target language. A source string $s_1^J = s_1 \dots s_j \dots s_J$ is to be translated into a target string $t_1^I = t_1 \dots t_i \dots t_I$. Among all possible target strings, the goal is to choose the string with the highest probability:

$$\tilde{t_1^I} = \underset{t_1^I}{\operatorname{argmax}}\, P(t_1^I | s_1^J)$$

where I and J are the number of words of the target and source sentence, respectively.

The first SMT systems were reformulated using Bayes' rule. In recent systems, such an approach has been expanded to a more general maximum entropy approach in which a log-linear combination of multiple feature functions is implemented [23]. This approach leads to maximising a linear combination of feature functions:

$$\tilde{t} = \underset{t}{\operatorname{argmax}} \left\{ \sum_{m=1}^{M} \lambda_m h_m(t, s) \right\}.$$

The overall architecture of this statistical translation approach is summarised in Figure 2.

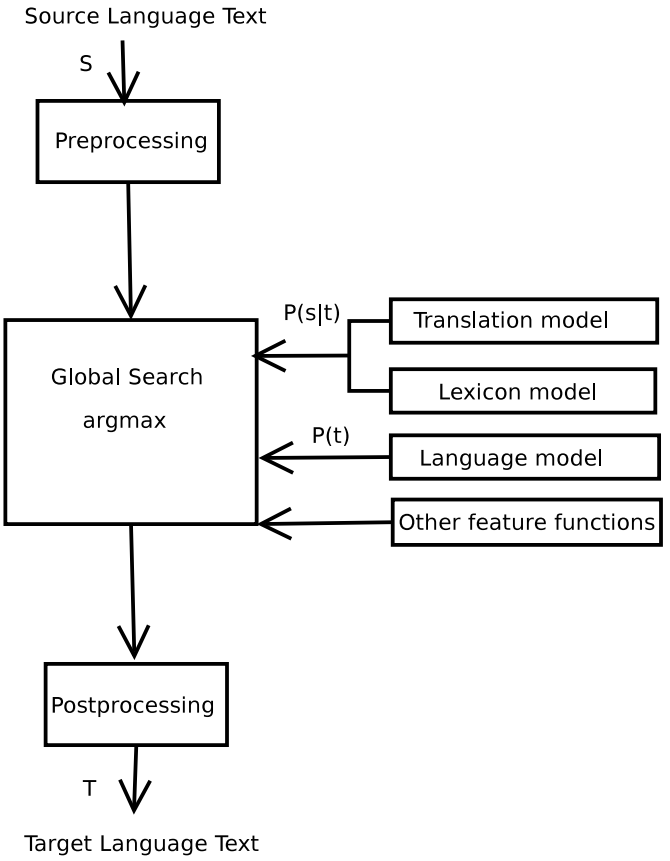

Fig. 2. Architecture of the SMT approach based on the log-linear framework approximation

The job of the translation model, given a target sentence and a foreign sentence, is to assign a probability that t_1^I generates s_1^J. While these probabilities can be estimated by thinking about how each individual word is translated, modern statistical MT is based on the intuition that a better way to compute these probabilities is by considering the behavior of phrases (sequences of words). The phrase-based statistical MT uses phrases as well as single words as the fundamental units of translation. Phrases are estimated from multiple segmentation of the aligned bilingual corpora by using relative frequencies.

The translation problem has also been approached from the finite-state perspective as the most natural way for integrating speech recognition and machine translation into a speech-to-speech translation system [29, 2, 6]. The Ngram-based system implements a translation model based on this finite-state perspective [11] which is used along with a log-linear combination of additional feature functions [19].

In addition to the translation model, SMT systems use the language model, which is usually formulated as a probability distribution over strings that attempts to reflect how likely a string occurs inside a language [7]. Statistical MT systems make use of the same n-gram language models as speech recognition and other applications do. The language model component is monolingual, so acquiring training data is relatively easy.

The lexical models allow the SMT systems to compute another probability to the translation units based on the probability of translating word per word of the unit. The probability estimated by lexical models tends to be less sparse in some situations than the probability given directly by the translation model. Many additional feature functions can also be introduced in the SMT framework to improve the translation, like the word or the phrase bonus.

3 STRUCTURAL COMPARISON OF RBMT AND SMT SYSTEMS

This section compares RBMT and SMT systems at the structural level. First, the different components and requirements of both systems are described; then, the challenges of each system are reported and, finally, their respective advantages and disadvantages are compared.

3.1 Components and Requirements

Building a rule-based MT system requires linguistic human knowledge and effort. In contrast, little written resources are needed. The main requirements of such systems are:

1. *A bilingual dictionary* that determines the words that can be translated.
2. *Linguistic rules* implemented by linguists, which require human effort and become more difficult as the pair of languages differ in their structure.

3. *Linguistic tools* like taggers and morphological analysers, to help the introduction of more generic rules. Additionally, they may be used to extend the bilingual dictionary in order to allow more words to be translated.

Building a statistical MT system requires written resources among other components. The main requirements are:

1. *Parallel corpus* at the sentence level. Generally speaking, less than 20 000 sentences produce unreadable translations. In case of having parallel corpus at the document level, there are some tools available to parallelise it at the sentence level.
2. *Training and translating software*, which is easy to obtain as it is available as open source:
   - Aligner: GIZA++ [24], BIA [17]. Given two parallel sentences, the aligner links at the level of words (generally, many-to-many alignments). Main statistic alignment models are easily explained in [15]. The word alignment is performed from source to target and target to source. Then, usually a symmetrization like the union is made.
   - Translation unit extractor (i.e. phrase [30] or tuple [19]).
   - Feature functions toolkits: SRILM [27], MOSES scripts. Each of the feature functions compute a probability for each phrase. These feature functions are combined in a loglinear framework presented earlier in this section.
   - Decoder: MOSES [16] , MARIE [9].

   Therefore, an SMT system theoretically does not require linguistic knowledge of the language pairs. In practice, linguistic knowledge may be important also for the statistically based systems, e.g. for the introduction of linguistic features that are to be observed.
3. High computational resources.

3.2 Challenges of RBMT and SMT

State-of-the-art rule-based MT approaches have the following challenges:

Semantic. RBMT approaches concentrate on a local translation. Usually, this translation tends to be literal and it lacks fluency. Additionally, words may have different meanings depending on their grammatical and semantic references.

Lexical. Words which are not included in the dictionary will have no translation. When keeping the system updated, new language words have to be introduced in the dictionary.

State-of-the-art statistical MT approaches have the following challenges:

Syntactic. Word ordering. For instance, SVO or VSO languages; location modifiers and nouns [28, 31, 8].

Morphological. Agreement. For instance, keeping number agreement when *Noun* + *Adjective* in Spanish [10, 22].

Lexical. Out-of-vocabulary words. Main reasons for out of vocabulary words are the limitation of training data, domain changes and morphology [18].

3.3 Advantages and Disadvantages of RBMT and SMT

Different approaches of MT have complementary pros and cons. Main core advantages and disadvantages of both approaches are shown in Table 1. At this point, no comparison is made at the level of performance, which will be studied later in the paper.

Advantages	Disadvantages
RBMT	
Based on linguistic theories	Requires linguistic rules and dictionaries
Adequate for languages with limited resources	Human Language Inconsistency (i.e. exceptions)
Does not require many computational resources	Disambiguation problems
Easy to perform error analysis	Local translations, Language dependent
	Expensive to maintain and extend
SMT	
No linguistic knowledge required	Requires parallel text
Reduces the human resources cost	Requires high computational resources
Easy to build	Difficult to perform error analysis
Easy to maintain (if data is available)	Problems languages pairs with different morphology/order
Trained with human translations	No linguistic background
Independent from the the pair of languages	

Table 1. Brief comparison of the advantages and disadvantages of the RBMT and SMT systems

It can be seen that both systems present strengths and weaknesses. Therefore, depending on the situation one system may be more adequate than the other. In addition, a combination of both may lead to a better and more efficient translation [20].

4 FREELY AVAILABLE CATALAN-SPANISH MT SYSTEMS: BRIEF DESCRIPTION

Some of the main Catalan-Spanish MT systems available on the web are introduced in this section. They include two RBMT systems: Apertium and Translendium,

and two SMT systems: Google and UPC. Next, a brief description of each system and its corporation is presented.

Apertium (`http://www.apertium.org/`)

The Apertium platform is an open-source machine translation system. Apertium was originally based on existing translators that had been designed by the Transducens group at the Universitat d'Alacant, and funded by the project Open-Source Machine Translation for the Languages of Spain. Subsequent development has been funded by the university, as well as by Prompsit Language Engineering.

The system uses a shallow-transfer machine translation methodology, initially designed for the translation between related language pairs. In addition to the translation engine, it also provides tools for manipulating linguistic data, and translators designed to run using the engine.

One of the main novelties of the Apertium architecture is that it has been released under open-source licenses (in most cases, GNU GPL; some data still have a Creative Commons license) and is distributed free of charge. This means that anyone having the necessary computational and linguistic skills will be able to adapt or enhance the platform or the language-pair data to create a new machine translation system, even for any pair of related languages. Apertium is designed according to the Unix philosophy: translation is performed in stages by a set of tools that operate on a simple text stream. Other tools can be added to the pipeline as required, and the text stream can be modified using standard tools.

Google (`http://translate.google.com/`)

Google's research group has developed its statistical translation system for the language pairs now available on Google Translate. Their system, in brief, feeds the computer with billions of words of text, both monolingual text in the target language, and aligned text consisting of examples of human translations between the languages. Then, they apply statistical learning techniques to build a translation model.

The *detect language* option automatically determines the language of the text the user is translating. The accuracy of the automatic language detection increases with the amount of text entered. Google is constantly working to support more languages to introduce them as soon as the automatic translation meets their standards. In order to develop new systems, they need large amounts of bilingual texts.

Translendium (`http://www.translendium.com`)

Translendium S.L., located in Barcelona, develops the Lucy Translator (LT) machine translation system, previously called Comprendium. Translendium is a Catalan company, subsidiary of the European group Lucy Software. The

Translendium team consists of linguists, lexicographers and computer scientists with more than 15 years experience in the machine translation field.

The system consists of a translation engine, with a modular structure of computational grammars and lexicons that makes possible to carry out a morphosyntactic analysis of the source text and then transfers it into the target language. This engine can be connected to translation memory modules and to a professional lexicon editor. In addition, it can be accessed through a multi-user task distribution server either from a web client or from a professional single user client. Moreover, the system can be adapted to general, social, technical and medical documents.

UPC (`http://www.n-ii.org/`)

This system has been developed at the Universitat Politècnica de Catalunya (UPC) and the work has been funded by the European Union under the integrated project TC-STAR: Technology and Corpora for Speech to Speech Translation (IST-2002-FP6-506738), by the Spanish Government under the project AVIVAVOZ: Methodologies for Speech-to-Speech Translation (TEC2006-13694-C03-01) and by the Catalan Government under the project TECNOPARLA.

The machine translation engine is based on an N-gram translation model integrated in an optimised log-linear combination of additional features. Thus the system is mainly statistical; however, a series of additional linguistic rules is included in order to solve some errors caused by the statistical translation, such as the ambiguity in adjective and possessive pronouns, orthographic errors or time expressions, among others.

Since time expressions differ largely in both languages, a detection-translation-generation module is added. The same procedure is used in the numbers, since many of them were not included in the training corpus. Other unknown words apart from numbers were solved by including a Spanish-Catalan dictionary as a post-process after the translation, and by a spell checker in order to avoid wrong-written – and thus unknown – words as input. The system is continuously updated by adding new corpora and the feedback of the users.

In the following sections, these systems are compared and they are all used with their respective versions date of *1st of April 2009*. Table 2 shows, for each system, the core methodology and limit of words that can be freely translated online.

	Core methodology	Limit of words
Apertium	Rule-based	no limit
Google	Statistical	no limit (if uploading files)
Translendium	Rule-based	2 500
UPC	Statistical	no limit (if uploading files)

Table 2. Comparison of the core methodology and limit of words to that can be freely translated online

5 EXPERIMENTAL FRAMEWORK

The aim of this section is to define an experimental framework in which the systems presented above can be compared both manually and automatically. The idea is to report the main differences in performance terms between the state-of-the-art rule-based and statistical systems. In the following sections, the results are reported through three different evaluation types: automatic, human and linguistic.

Two test sets are defined in order to perform the evaluation. The first one is a compilation of journalistic material. The Spanish source test corpus consists of 711 sentences extracted from *El País* and *La Vanguardia* newspapers, and the Catalan source test corpus consists of 813 sentences extracted from the *Avui* newspaper and transcriptions from the TV program *Àgora*. For each set and each direction of translation, two manual references are provided. Table 3 shows the number of sentences, words and vocabulary used for each language.

	Spanish	Catalan
Sentences	711	813
Words	15 974	17 099
Vocabulary	5 702	5 540

Table 3. Corpus statistics for the journalistic Catalan-Spanish test set

A second test corpus is provided within the medicine domain. This medical corpus was kindly provided by the UniversalDoctor company, which focuses on facilitating communication between healthcare providers and patients from various origins (`http://www.universaldoctor.com`). The medical corpus consists of 554 parallel sentences and only one manual reference for each direction of translation was available. Table 4 shows the number of sentences, words and vocabulary used for each language.

	Spanish	Catalan
Sentences	554	554
Words	3 127	3 117
Vocabulary	920	913

Table 4. Corpus statistics for the medical Catalan-Spanish test set

6 AUTOMATIC EVALUATION

Automatic evaluation is one of the most crucial issues in the development stage of a MT system, given that other types of evaluation are usually expensive. Error rate is typically measured by comparing the system output against a set of human references, according to an evaluation metric at choice. By far, the most widely used metric in recent literature is BLEU (Bilingual Evaluation Understudy). It is

a quality metric and it is defined in the range between 0 and 1 (or in percentage between 0 and 100), 0 meaning the worst translation (where the translation does not match the reference in any word), and 1 the perfect translation. BLEU computes lexical matching accumulated precision for n-grams up to length four [25]. Apart from BLEU, in this paper other automatic evaluation measures are used to compare hypothesis translations against supplied references that have been introduced.

NIST [12] is an accuracy measure that calculates how informative a particular n-gram is, and the rarer a correct n-gram is, the more weight it will be given. Small variations in translation length do not impact much in the overall score. On the other hand, Translation Error Rate (TER) [26] is an error metric that measures the number of edits required to change a system output into one of the references.

Word Error Rate (WER) [21] is a standard speech recognition evaluation metric. A general difficulty of measuring performance lies in the fact that the translated word sequence can have a different length from the reference word sequence (supposedly the correct one). The WER is derived from the Levenshtein distance, working at the word level. Similar to WER, the Position-Independent Error Rate (PER) is again computed on a sentence-by-sentence basis. The main difference with WER is that it does not penalise the wrong order in the translation.

Tables 5 and 6 show the results obtained through automatic evaluations in the journalistic and medical corpora, respectively. It can be clearly seen that, in the journalistic translations, statistical systems perform better in terms of automatic evaluation than rule-based systems. These results may be explained with the fact that statistical systems use journalistic corpora to train their systems, since this kind of corpora can be easily collected.

	Es2Ca				
	BLEU	NIST	TER	WER	PER
Apertium	84.24	13.17	11.75	11.83	11.12
Google	86.10	13.33	11.32	11.22	10.47
Transl	85.97	13.35	11.04	11.14	10.43
UPC	86.54	13.40	10.76	10.88	10.08

	Ca2Es				
	BLEU	NIST	TER	WER	PER
Apertium	86.21	13.24	8.55	8.64	8.05
Google	92.37	13.81	5.70	5.81	5.33
Transl	87.81	13.37	7.83	7.92	7.30
UPC	88.58	13.46	7.80	7.91	7.27

Table 5. Automatic evaluation using the journalistic corpus

In the Catalan-to-Spanish medical translation results, Translendium provides the best translation. As explained in Section 4, Translendium has the option of translating medical documents. This means that the system has been particularly

	Es2Ca				
	BLEU	NIST	TER	PER	WER
Apertium	54.15	7.98	34.64	35.10	32.52
Google	55.05	8.06	33.93	34.58	31.87
Transl	57.32	8.36	30.97	31.59	28.98
UPC	56.66	8.29	32.18	32.80	30.12

	Ca2Es				
	BLEU	NIST	TER	PER	WER
Apertium	50.92	7.66	37.19	37.78	34.88
Google	57.09	8.28	32.61	33.13	30.17
Transl	54.44	8.05	34.04	34.57	31.60
UPC	55.34	8.11	34.26	34.75	31.66

Table 6. Automatic evaluation using the medical corpus

built for this semantic domain. However, in the Catalan-to-Spanish medical translation, Google offers the best system, since it might be using use a big target language model (see Section 2) that helps to obtain a more fluent translation.

Although results show coherence in all automatic measures – which gives more consistency to the evaluation – all measures present several deficiencies that cast serious doubts on the coherence with human criteria and on its usefulness, both for sentence-level error analysis and for system-level comparison [5]. Moreover, the automatic measures do not give any information about the type of errors committed by the systems. In order to perform a more objective and accurate evaluation, the following sections present a human evaluation and a proposal for a linguistic error evaluation.

7 HUMAN EVALUATION

The comparison between different translation system outputs was performed by 10 different human evaluators. All the evaluators were bilingual in Catalan and Spanish, therefore, no reference of translation was shown to them, in order to avoid any bias in their evaluation.

The evaluators performed two types of comparisons. First, each evaluator was asked to compare all the systems. Figure 3 shows an example of the screenshot that is shown to the annotator. Each evaluator compared 100 randomly extracted translation pairs, and assessed in each case whether one system produced a better translation, or whether two or more were equivalent. Each judge evaluated a different set of (possible overlapping) sentences. In order to avoid any bias in the evaluation, the respective position in the display of the sentences corresponding to each system was also random. The comparison among the four systems was not taken into account when two systems or three systems were better than the other(s) one(s). It counts strictly when one system is better than the others. Making the 100 sentence-

judgements on the four systems usually takes one hour. We collected a total of 1 000 judgements in the comparison of the 4 systems. Although the evaluation task is different, the number of items that were judged for our task is comparable to the judgements collected in the evaluation of English-Spanish WMT 2009 task (the yes/no task included 878 items judged) [4]. Therefore, results present a statistical significance equivalence similar to the ones in that international evaluation. Tables 7 and 8 show average results in percentage for the journalistic corpus and for the medical corpus, respectively.

LINE 39

Source: Cal que hi hagi oferta per a tothom.
(1): Es necesario que haya oferta para todos.
(2): Hace falta que haya ofrecida para todo el mundo.
(3): Hace falta que haya oferta para todo el mundo.
(4): Es preciso que haya oferta para todos.
Which translation was better (0 for same quality)?

Fig. 3. Screenshot of the human evaluation when comparing the four systems

Es2Ca				Ca2Es			
Apertium	Google	Transl	UPC	Apertium	Google	Transl	UPC
10 %	19 %	30 %	**41 %**	9.0 %	36 %	**37 %**	18 %

Table 7. Human judgements regarding comparison of all four systems using the journalistic corpus. Each column indicates the number of times (in percentage) in which one system was chosen as better than the others

Es2Ca				Ca2Es			
Apertium	Google	Transl	UPC	Apertium	Google	Transl	UPC
11 %	11 %	**55 %**	22 %	7 %	21 %	**57 %**	15 %

Table 8. Human judgements regarding comparison of all four systems using the medical corpus. Each column indicates the number of times (in percentage) in which one system was chosen as better than the others

Second, each judge was asked to make a system-to-system (pairwise) comparison. Each annotator evaluated 100 randomly extracted translation pairs, and assessed in each case whether one system produced a better translation than the other one, or whether the two outputs were equivalent. Figure 4 shows an example of the screenshot shown to the annotator. Each judge did this evaluation for three pairs of systems. Therefore, a total number of 3 000 judgements was collected, i.e. 300 judgements of each pair of systems (3 different evaluators for each pair of systems, evaluating 100 sentences). Tables 5 and 6 show average results in percentage for the journalistic corpus and for the medical corpus, respectively.

LINE 39

Source: Cal que hi hagi oferta per a tothom.
(1): Hace falta que haya ofrecida para todo el mundo.
(2): Es necesario que haya oferta para todos.
Which translation was better (0 for same quality)?

Fig. 4. Screenshot of the human evaluation when comparing system-to-system

Es2Ca	Google	Transl	UPC
Apertium	48 %	39 %	38 %
	Google	40 %	46 %
		Transl	48 %
Ca2Es	Google	Transl	UPC
Apertium	43 %	30 %	37 %
	Google	58 %	51 %
		Transl	56 %

Fig. 5. Human judgements after the system-to-system comparison using the journalistic corpus. Results show in which percentage the systems in the left column where marked as better than the systems in the upper row.

The results show coherence among both human evaluations:

1. comparison of the four systems and
2. system-to-system comparison.

In all medical translations, Translendium shows the best performance, while in the Catalan-to-Spanish journalistic translations the best performance is provided by UPC system. In the opposite direction, the best system is Translendium again.

Given the two types of human evaluation, the system-to-system comparison gives more polarised results than when comparing all systems. This is due to the

Es2Ca	Google	Transl	UPC
Apertium	48 %	27 %	42 %
	Google	28 %	48 %
		Transl	77 %
Ca2Es	Google	Transl	UPC
Apertium	41 %	21 %	45 %
	Google	24 %	53 %
		Transl	84 %

Fig. 6. Human judgements after the system-to-system comparison using the medical corpus. Results show in which percentage the systems in the left column where marked as better than the systems in the upper row.

definition of each evaluation. Comparing four systems leaves out those sentences where two or three systems have the same outputs, whereas comparing two systems leaves out those sentences only if the quality of both systems is equivalent, which occurs less times than in the comparison of four systems.

When comparing automatic measures and human judgements, results tend to be different (except in the Spanish-to-Catalan journalistic task). For instance, in the case of the Catalan-to-Spanish direction, Translendium provides better results in the human judgements instead of Google. The same happens in the Spanish-to-Catalan medical task. Human judgements correlate better with automatic evaluation when comparing systems of the same core methodology.

8 LINGUISTIC EVALUATION

In order to evaluate and compare several machine translation systems in a proper way, it becomes necessary to perform both automatic and human evaluations. However, it may be useful to perform an additional evaluation according to some linguistic criteria. To this end, this paper evaluates the selected machine translation systems according to the kind of errors they make in the translation task.

Next, a linguistic error classification is proposed in order to linguistically evaluate the encountered errors. Then, an evaluation is presented according to the proposed classification.

8.1 Linguistic Error Classification

The errors are reported according to the following linguistic levels involved: orthographic, morphological, lexical, semantic and syntactic. A detailed description of which errors are included in these levels is needed for a proper error analysis. To this end, an error classification is next described according to the specific cases that can be found in a Catalan to Spanish (and vice versa) translation task.

8.1.1 Orthographic Errors

Orthography refers to the correct way of using a specific writing system to write a language. In the Catalan-Spanish pair, the most common errors are the following:

- Punctuation marks
  - *Exclamation and interrogation marks*: unlike Catalan, Spanish uses these marks at both the beginning and the end of the sentences; therefore, it is common to find a missing or wrong-placed mark in Spanish or a spare mark in Catalan.
  - *Other punctuation marks*: full stops, commas, colons, semicolons, dots, etc.: it is also frequent to find some errors related to these punctuation marks: commas instead of semicolons, missing full stops or dots, full stops instead of colons, etc.

- Accents
  - *Accented vowels when not necessary*
  - *Missing accents*
  - *Erroneous accents*
- Capital and lower case letters
  - *Capital letters within a sentence*
  - *Lower case letters at the beginning of a sentence*
  - *Lower case letters in acronyms or proper nouns*
- Joined words: A less common error involves joining two consecutive words (e.g. $y\ a \rightarrow ya$).
- Spare blanks: After translation itself, detokenisation is a complex process that requires many rules and exceptions. It is not easy to know whether quotes, for instance, should be joined to the word placed in the next position or not. This leads to spare blanks due to a non-detokenisation when required or to a detokenisation in the wrong direction.
- Apostrophe: In Catalan, some articles and prepositions are – with some exceptions – apostrophized in front of vowels. This difference with Spanish leads to apostrophe errors: missing apostrophes in Catalan or spare apostrophes in Spanish.
- Conjunctions: Conjunctions *i* and *o* in Catalan are translated into *y* and *o*, except for the cases in which the following word begins with an *i* or *u*, respectively. These differences lead to orthographic errors in both directions of translation.
- Foreign words: Translating foreign words becomes always a complex task that produces diverse errors, such as the use of erroneous accents, among others.

8.1.2 Morphological Errors

Morphology refers to identification, analysis and description of the structure of words. Within this framework, the following morphological types of errors are encountered:

- Lack of gender concordance.
- Lack of number concordance.
- Apocopes: In Spanish, some adjectives are shortened when they precede a noun. They are called *apocopes*, and if the translator is not able to realize such structure change, a morphological error is then encountered.
- Verbal morphology: A verbal morphology error refers to a verb that is not well inflected. The most common inflection errors are translation of an inflected verb into the infinitive form, or lack of person concordance.

- Lexical morphology: Lexical morphology concerns basically word formation: derivation and compounding. A common error of this type is the use of a derivate in a wrong way (e.g. *liguero* instead of *de la Liga*) or a wrong compounding (e.g. *històric-social* instead of *historicosocial*).
- Morphosyntactic changes: Morphological changes due to syntactic structure can occur (e.g. *a mi* (to me) but *con mi* (with me) results in the form *conmigo*).

8.1.3 Lexical Errors

Lexis refers to the total bank of words and phrases of a particular language. The most common errors found in the translation process are the following.

- No correspondence at all between the source and target words.
- Source word not translated and left intact.
- Source word not translated, missing target word.
- Non-translated proper nouns or translated when not necessary.

8.1.4 Semantic Errors

Semantics refers to the meaning of the words. Frequently, the same source word has two or more possible meanings. When the non-correct meaning is chosen, a semantic error is encountered. Semantic errors can be related to polysemy or homonymy.

- Polysemy
- Homonymy
- Expressions used in a different way.

8.1.5 Syntactic Errors

Syntax refers to the principles and rules for constructing sentences in natural language. Common errors that can be found in this linguistic level are the listed next.

- Prepositions
  - *Lack of preposition a in front of a target Spanish direct object.*
  - *Preposition not elided in the target language.*
  - *Preposition not inserted in the target language.*
  - *Source preposition maintained, instead of a new correct target preposition.*
- Relative clauses.
- Verbal periphrasis.
- Clitics: incorrect syntactic function (*leísmo* and *laísmo*) or wrong clitic-verb combination.
- Missing or spare article in front of proper nouns.
- Syntactic reordering of the elements of the sentence.

8.2 Evaluation and Results

After having defined the specific error corresponding to each linguistic error, all the translations were manually analysed in order to detect the number of errors. Tables 9 and 10 show the number of errors detected in each translation system for the Es2Ca and Ca2Es directions, respectively. More specifically, both tables show the test sentences that contained at least one error, the total number of errors, and the number corresponding to each specific type of error. The best results are shown in bold numbers.

Translator	Sentences with errors	Total errors	Orthogr.	Morphol.	Lexical	Seman.	Syntac.
Apertium	380	608	**14**	28	146	200	220
Google	378	646	169	80	113	101	183
Transl	272	399	26	**20**	68	153	132
UPC	**229**	**296**	62	29	**65**	**61**	**79**

Table 9. Number and type of errors encountered in the 711-sentence Es2Ca translations

Translator	Sentences with errors	Total errors	Orthogr.	Morphol.	Lexical	Seman.	Syntac.
Apertium	414	593	63	29	99	237	165
Google	265	357	102	40	54	**50**	111
Transl	274	364	**58**	**24**	**47**	125	110
UPC	**258**	**338**	82	37	67	65	**87**

Table 10. Number and type of errors encountered in the 813-sentence Ca2Es translations

Some considerations can be extracted from these tables. First of all, the best performance is achieved in the UPC system for both directions. These results are correlated with the automatic ones in Table 5 in the Es2Ca direction, but not in the opposite one. Thus, it seems that although Google committed more errors in the Ca2Es translation in absolute values, these were much less important in automatic evaluation terms.

Second, the results also show a correlation according to the type of errors encountered. While the RBMT systems (Apertium and Translendium) committed less orthographic and morphological errors, the SMT systems (Google and UPC) performed better at the semantic level. In contrast, the performance at the other levels (lexical and syntactic) did not follow this classification.

In addition to these results, it becomes of interest to have an idea of the most common errors found in each linguistic level. To this end, a 50 % of the errors found were randomly selected in order to manually analyse and compute how many specific subtypes of errors contained. In Figure 7, a graphic for each linguistic level is plotted, in which the percentages of the error sublevels are represented. At the orthographic level, punctuation marks provided most of the errors. Lack of gender

concordance was the most common morphological error, not translated source word was the major lexical error, polysemy and homonymy were the most encountered semantic errors, and prepositions were the main syntactic problem.

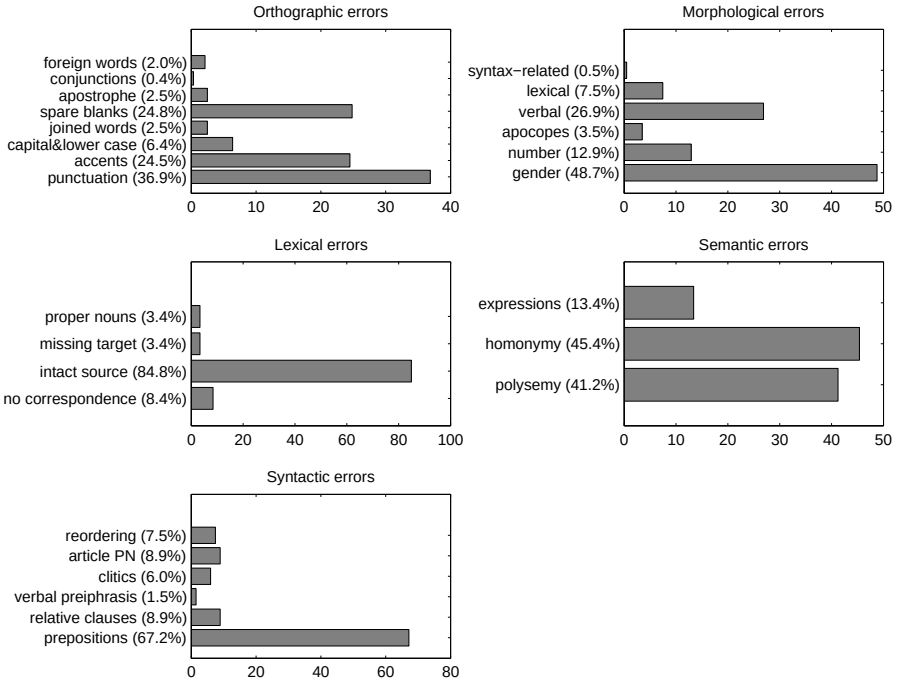

Fig. 7. Percentage of the error sublevels

As in the previous evaluations, a further analysis was performed within the medicine domain. To this end, the medical test corpus consisting of 554 parallel Catalan-Spanish sentences was used. Tables 11 and 12 show the total number of errors encountered and the number of errors corresponding to each of the linguistic levels and for each direction of translation.

Translator	Sentences with errors	Total errors	Orthogr.	Morphol.	Lexical	Seman.	Syntac.
Apertium	118	132	**0**	6	23	46	57
Google	137	149	5	23	46	**26**	49
Transl	**74**	**80**	**0**	**1**	**9**	29	**41**
UPC	109	123	5	8	28	32	50

Table 11. Number and type of errors encountered in the 554-sentence Es2Ca translations using the medical corpus

Translator	Sentences with errors	Total errors	Orthogr.	Morphol.	Lexical	Seman.	Syntac.
Apertium	191	239	33	10	31	133	32
Google	131	164	36	19	37	**58**	**14**
Transl	**105**	**105**	**3**	**4**	**6**	68	24
UPC	165	189	35	10	31	72	41

Table 12. Number and type of errors encountered in the 554-sentence Ca2Es translations using the medical corpus

Again, a correlation with the automatic evaluation is observed in the Es2Ca translation. Since in the Translendium system an option for the medical domain is provided, less errors are committed in the translation. In contrast, automatic evaluation values are higher for the Google system in the opposite direction. It seems, thus, that although much more errors are encountered in Google, these are not influential enough to give worse automatic evaluation results.

A correlation is also observed between the type of errors and the core methodology of the systems. As in the journalistic domain experiments, orthographic and morphological errors tend to be lower in the RBMT systems Apertium and Translendium. The best lexical and syntactic achievements are also found in both kinds of systems. However, the best semantic performance is achieved by Google and Translendium instead of Google and UPC. The difference lies again in the fact that Translendium is optimised for a medical domain, so that many semantic problems derived from the context meaning are avoided.

8.3 Comparison of Human and Linguistic Evaluations

By analysing the results obtained in the linguistic evaluation, some correlations have been observed with the automatic evaluation results. In this section, the same analysis is performed in order to see what human and linguistic evaluations have in common and whether they are characterised by any sort of correlation.

The evaluation comparison is performed as follows. When a system S1 was selected as better than a system S2, we checked in which levels S1 got less errors than S2. Thus, it was found that, when a system was better than another one, it was also better at the orthographic level in 42 % of the cases, at the morphological level in 50 % of the cases, at the lexical level in 75 % of the cases, and in both semantic and syntactic levels in 83 % of the cases.

These results led us to two considerations: in the first place, humans rely most on syntactic and semantic errors than in orthographic, morphological errors and lexical errors. Second, the percentage of cases is directly proportional to the depth of the language structure: orthographic level is related to the most superficial structure, while syntactic level is related to the deepest one, so that human evaluators use to penalise strongly those errors committed at deeper levels.

9 CONCLUSIONS

The aim of this work was to analyse the main differences between rule-based and statistical machine translation paradigms in the specific case of Catalan-Spanish pair. The perfect paradigm does not exist, as neither does the perfect evaluation method. In this paper, three different kinds of evaluation have been carried out in order to compare them and to find common characteristics, especially between the proposed linguistic evaluation.

A correlation has been found between the type of linguistic errors committed and the core methodology of the systems. Orthographic and morphological errors tend to be lower in the rule-based machine translation systems, while the performance at the semantic level is better in the statistical systems – because of the context given by the training corpus – except in the rule-based systems optimised for the semantic domain it is been treated, as in the case of Translendium which can be optimised for a medical domain. Further system combination research work may take these results into account. The system combination may take advantage of rule-based systems regarding orthographic and morphological aspects of the translation, and statistical systems regarding semantic aspects of the translation.

Furthermore, correlation with the automatic evaluation was also found, as well as with the human evaluation. It can be inferred from the results that human evaluators tend to penalise the systems much stronger when the errors committed come from deeper language structures.

In the future, we plan to experiment with more language pairs and to see if the results differ when working with two typologically different languages.

Acknowledgements

This work has been partially funded by the Spanish Ministry of Economy and Competivity through the Juan de la Cierva fellowship program. The authors also want to thank the Barcelona Media Innovation Centre for its support and permission to publish this research.

REFERENCES

[1] Arnold, D.—Balkan, L.: Machine Translation: An Introductory Guide. Comput. Linguist., Vol. 21, 1995, No. 4, pp. 577–578.

[2] Bangalore, S.—Riccardi, G.: Finite-State Models for Lexical Reordering in Spoken Language Translation. In: Proc. of the 6th Int. Conf. on Spoken Language Processing, ICSLP '02, Vol. 4, Beijing 2000.

[3] Berger, A. L.—Brown, P. F.—Della Pietra, S. A.– Della Pietra, V. J.—Gillett, J. R.—Lafferty, J. D.—Mercer, R. L.—Printz, H.—Ureš, L.: The Candide System for Machine Translation. In: HLT '94: Proceedings of the Workshop on Human Language Technology 1994.

[4] Callison-Burch, C.—Koehn, P.—Monz, C.—Schroeder, J.: Findings of the 2009 Workshop on Statistical Machine Translation. In: Proceedings of the Fourth Workshop on Statistical Machine Translation, Association for Computational Linguistics, Athens, Greece 2009.

[5] Callison-Burch, C.—Osborne, M.—Koehn, P.: Re-Evaluating the Role of BLEU in Machine Translation Research. In: Proc. of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, Italy 2006.

[6] Casacuberta, F.: Finite-State Transducers for Speech-Input Translation. In: IEEE Automatic Speech Recognition and Understanding Workshop, ASRU, Trento 2001.

[7] Chen, S. F.—Goodman, J. T.: An Empirical Study of Smoothing Techniques for Language Modeling. Technical report, Harvard University 1998.

[8] Costa-Jussà, M.—Fonollosa, J.: An n-Gram-Based reordering model. Computer Speech and Language, Vol. 23, 2009, No. 3, pp. 362–375.

[9] Crego, J. M.: Architecture and Modeling for n-Gram-Based Statistical Machine Translation. Ph. D. thesis, Department of Signal Theory and Communications, Universitat Politècnica de Catalunya (UPC) 2008.

[10] De Gispert, A.—Gupta, D.—Popovic, M.—Lambert, P.—Mario, J.—Federico, M.—Ney, H.—Banchs, R.: Poving Statistical Word Alignments with Morphosyntactic Transformations. Proc. of 5th Int. Conf. on Natural Language Processing (FinTAL '06), Vol. 6, pp. 368–379.

[11] De Gispert, A.—Mariño, J.: Using x-Grams for Speech-to-Speech Translation. In: Proc. of the 7th Int. Conf. on Spoken Language Processing, ICSLP '02, Denver 2002.

[12] Doddington, G.: Automatic Evaluation of Machine Translation Quality Using nGram Co-Occurrence Statistics. In: Proc. of the Human Language Technology Conference, HLT-NAACL '02, San Diego 2002.

[13] Dorr, B. J.: Machine Translation Divergences. Computational Linguistics, Vol. 20, 1994, No. 4, pp. 597–633.

[14] González, G. A.—Boleda, G.—Melero, M.—Badia, T.: Traduccin Automática Estadística Basada en n-Gramas. Procesamiento del Lenguaje Natural, SELPN, Vol. 35, 2005, pp. 69–76.

[15] Knight, K.: A Statistical Machine Translation Tutorial Workbook. `http://www.clsp.jhu.edu/ws99/projects/mt/wkbk.rtf` (April 1999).

[16] Koehn, P.—Hoang, H.—Birch, A.—Callison-Burch, C.—Federico, M.—Bertoldi, N.—Cowan, B.—Shen, W.—Moran, C.—Zens, R.—Dyer, C.—Bojar, O.—Constantin, A.—Herbst, E.: Moses: Open Source Toolkit for Statistical Machine Translation. In: Proc. of the 45th Annual Meeting of the Association for Computational Linguistics, Prague (Czech Republic) 2007.

[17] Lambert, P.—Banchs, R. E.—Crego, J.: Discriminative Alignment Training Without Annotated Data for Machine Translation. In: Proc. of the Human Language Technology Conference, HLT-NAACL '07, Rochester (USA) 2007.

[18] Langlais, P.—Patry, A.: Translating Unknown Words by Analogical Learning. In: Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL) 2007.

[19] Mariño, J.—Banchs, R.—Crego, J.—De Gispert, A.—Lambert, P.—Fonollosa, J.—Costa-Jussà, M.: N-Gram Based Machine Translation. Computational Linguistics, Vol. 32, 2006, No. 4, pp. 527–549.

[20] Matusov, E.—Leusch, G.—Banchs, R. E.—Bertoldi, N.—Dechelotte, D.—Federico, M.—Kolss, M.—Lee, Y.—Marino, J. B.—Paulik, M.—Roukos, S.—Schwenk, H.—Ney, H.: System Combination for Machine Translation of Spoken and Written Language. IEEE Transactions on Audio, Speech and Language Processing, Vol. 16, 2008, No. 7, pp. 1222–1237.

[21] McCowan, I. M.—Moore, D.—Dines, J.—Gatica-Perez, D.—Flynn, M.—Wellner, P.—Bourlard, H.: On the Use of Information Retrieval Measures for Speech Recognition Evaluation. IDIAP-RR 73, IDIAP, Martigny, Switzerland 2004.

[22] Niessen, S.—Ney, H.: Statistical Machine Translation With Scarce Resources Using Morpho-Syntactic Information. Computational Linguistics, Vol. 30, 2004, No. 2, pp. 181–204.

[23] Och, F.: Minimum Error Rate Training in Statistical Machine Translation. In: Proc. of the 41th Annual Meeting of the Association for Computational Linguistics, Sapporo 2003.

[24] Och, F.—Ney, H.: A Systematic Comparison of Various Statistical Alignment Models. Computational Linguistics, Vol. 29, 2003, No. 1, pp. 19–51.

[25] Papineni, K.—Roukos, S.—Ward, T.—Zhu, W.-J.: Bleu: A Method for Automatic Evaluation of Machine Translation. In: Proc. of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA 2002.

[26] Snover, M.—Dorr, B. E.—Schwartz, R.—Micciulla, L.—Makhou, J.: A Study of Translation Edit Rate With Targeted Human Annotation. In: Proc. of the 5th Conf. of the Association for Machine Translation in the Americas (AMTA '06), Boston (USA) 2006.

[27] Stolcke, A.: SRILM – An Extensible Language Modeling Toolkit. In: Proc. of the 7th Int. Conf. on Spoken Language Processing (ICSLP '02), Denver (USA) 2002.

[28] Tillmann, C.—Ney, H.: Word Reordering and a Dynamic Programming Beam Search Algorithm for Statistical Machine Translation. Computational Linguistics, Vol. 29, 2003, No. 1, pp. 97–133.

[29] Vidal, E.: Finite-State Speech-to-Speech Translation. In: Proc. Int. Conf. on Acoustics Speech and Signal Processing, Munich 1997.

[30] Zens, R.—Och, F.—Ney, H.: Phrase-Based Statistical Machine Translation. In: M. Jarke, J. Koehler, G. Lakemeyer (Eds.), KI – 2002: Advances in artificial intelligence, LNAI Vol. 2479, Springer Verlag 2002, pp. 18–32.

[31] Zhang, Y.—Zens, R.—Ney, H.: Chunk-Level Reordering of Source Language Sentences With Automatically Learned Rules for Statistical Machine Translation. In: Proc. of the Human Language Technology Conf. (HLT-NAACL '06): Proc. of the Workshop on Syntax and Structure in Statistical Translation (SSST), Rochester 2007.

Marta R. Costa-Jussà received her Ph. D. in statistical machine translation from the Universitat politcnica de Catalunya, UPC, Spain in 2008. She has done research at LIMSI-CNRS (Paris) and (Singapore). Currently, she works in Barcelona Media Innovation Center under a "Juan de la Cierva" research fellowship program. She has participated in more than 10 European and Spanish research projects and published over 70 papers in international journals and conferences. She combines her research work together with university teaching and consultancy in translation companies.

Mireia Farrús graduated in physics and linguistics from the University of Barcelona and received her Ph. D. from the Technical University of Catalonia. She has done research in Saarland University (Germany), Umeå University (Sweden) and University of Canberra (Australia) in the speech and language field. She has recently joined the Pompeu Fabra University (Spain) as a Visiting Professor, where she works in e-learning applications related to natural language processing.